A Flexible Method of Customer Activities Recognition in Retail Store

修士学位論文要約（令和２年３月）

店頭における消費者行動の柔軟な認識手法

温　佳昊
指導教員：菅沼　拓夫,　学位論文指導教員：阿部　亨

A Flexible Method of Customer Activities Recognition in Retail Store

Jiahao WEN
Supervisor: Takuo SUGANUMA,　Research Advisor: Toru ABE

Customer Activities (CA) in retail environments, which refer to the interaction of customers with products and services, provide valuable information for marketing. There are different levels and types of CA, and required CA vary according to marketing environments. Many methods of Customer Activity Recognition (CAR) from videos have been proposed, and each of them is mostly designed to recognize specified CA by using end-to-end (e2e) models. Consequently, such existing methods cannot easily modify and extend CAR due to needs of rebuilding different e2e models entirely. We propose a flexible CAR method, where CA are organized into a hierarchy with several different levels, and different types of CA are recognized for each level from CAR results for lower levels. Since the proposed method preforms CAR separately for each level, CAR can be updated individually for each level, which means that CAR can be easily modified and extended depending on the changes in marketing environments.

1. Introduction

The innovation of traditional retail is necessary to survive the competition with online shopping. As shown in Fig. 1, traditional retail environments only use purchase records to analyze the purchasing behaviors to support the marketing plan[1]. However, records only show the final results of purchase decisions without the process of making those decisions. As the process probably reveals the reasons for those decisions, it is also valuable information for the marketing plan. Therefore, various sensors, especially the cameras, are used for collecting real-time data to reveal the shopping process. This solution is called “smart retail.” With machine learning models, smart retail is able to deal with those real-time data to get information about the shopping process. In such information, we focus on Customer Activities (CA) in retail environments, which refer to the interaction of customers with products and services, including customer’s position, movement, behavior, etc.

2. Problems

Many methods of Customer Activity Recognition (CAR) have been proposed[2]. They track objects in retail environments and tracking results are mostly utilized for customer behavior recognition. Those existing methods utilized a machine learning-based end-to-end (e2e) model which cause three problems.

- First and foremost, an e2e model outputs the same type of data. Since marketing requires different types of data, many e2e models are necessary to get various types of data. Therefore, it spends lots of time and computational resources when training and using those e2e models.
- Besides, it is hard to modify the model’s outputs. The changes of marketing demands require modification of the model’s output to recognize new behaviors. However, the modification, namely the retraining of an e2e model is time-consuming.
- Last but not least, the e2e model cannot be partially updated. Because the e2e model is a black box, it is impossible to figure out the people detection part in the behavior recognition model and update only this part by the latest method.

Those three problems make the existing methods inflexible. “inflexible” means the existing methods cannot be easily modified to adapt to the changes of marketing demands and retail environments. Additionally, none of them mentioned the reason why they would like to recognize those behaviors. Therefore, it is hard to say those methods can provide valuable information for marketing.

3. Proposal

To solve those problems, we propose a multi-level hierarchy to organize CAR as shown in Fig. 1. We investigated the outputs of existing CAR methods and find that they have a similar flow on CAR. They detect the objects’ position in each frame (Level 1 “Position”), then track and analyze the movement of each object (Lv.2 “Movement”), finally recognize behavior

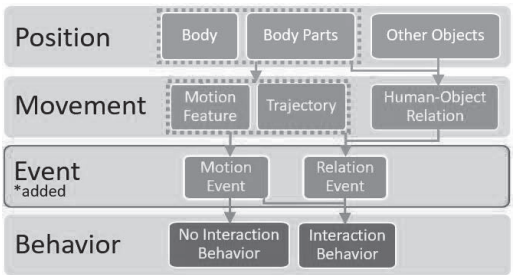

Fig. 1. Proposed Hierarchy

from the tracking outputs (Lv.4 “Behavior”).

Human behavior is defined as a composition of multiple events and an event refers to a single entity that cannot be decomposed[3]. It reveals that we recognize events primarily and then combine them together to be the behavior. Therefore, a new level “Event” is added and defined as above. The Lv.3 “Event” reduces the input trajectory from “Movement” to compress some repeated data. Then, the reduced trajectories are symbolized to avoid the usage of any numerical value and make it more understandable.

With “Event”, the process of behavior recognition becomes searching the event combination instead of the dataflow through a black box. The recognition of behaviors can be easily modified. Also, different levels in the hierarchy output different types of data, and the update of any level has no influence on the other levels. Thus, the proposed hierarchy is flexible.

4. Evaluation

As the method in each level can be easily modified, evaluating the hierarchy by each level’s accuracy makes no sense. Therefore, we reveal the relation between the Accuracy (Acc.) of Levels (Lv.) 1, 2 and the Acc. of Lv.3, 4. Hence, we modify Lv.1, 2 Acc. and observe changes of Lv.3, 4 Acc.

The camera is installed as shown in Fig. 2 to get the top-view image. Those participants are requested to pick at least one product with hands from the shelf. The person, hand, bottle, spray, wet-tissue, and pear-water totally six objects are used in the experiment. And seven behaviors are defined and recognized by the implemented hierarchy. We manually

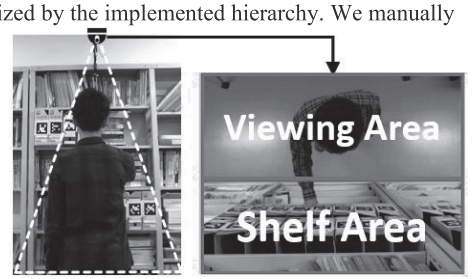

Fig. 2. Camera Installation and Input Image

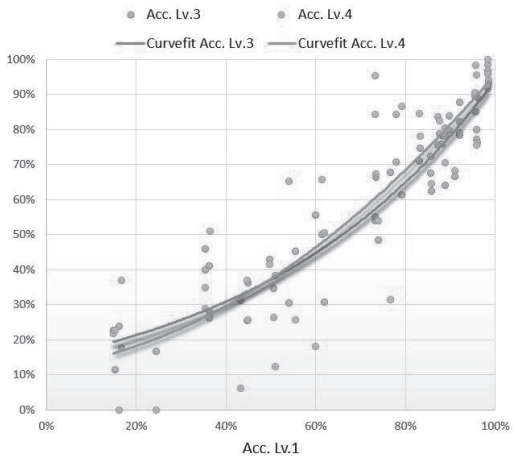

Fig. 3. Relation between Acc. Lv.1 & Acc. Lv.3, 4

labeled and modified the ground truth (GT) of object position for each video. Then, input those modified positions into the implemented hierarchy to get the results of Lv.3 and Lv.4. Hence, input the GT of Lv.1, 2, the model’s results are the GT of Lv.3, 4.

The Acc. of Lv.1, 2 is the average intersection over union (IOU) of all objects in all frames of a video. The Acc. of Lv.3, 4 is their F1-measure of a whole video. Totally 2462 frames are labeled which contain 744 events and 121 behaviors in GT. The result is as shown in Fig. 3. The curve-fit reveals that the Acc. Lv.3, 4 and the Acc. Lv.1 has a positive correlation. The increase of Acc. Lv.1 will increase the Acc. of higher levels. With experiment results, to indicate that the proposed hierarchy and method can recognize not only behaviors in existing methods but also the behavior (selecting) not in existing methods, we defined all behaviors in existing methods by event sequence. Therefore, it indicates that the proposed model is at least more flexible than existing methods.

5. Conclusion

To adapt to the dynamic market demands in smart retail, we proposed a multi-level hierarchy to organize CAR well and a flexible method to utilize the hierarchy. With the evaluation results, it can be concluded that our proposed hierarchy and methods are flexible.

Reference

1) J. Liu et al., “Customer Behavior Recognition in Retail Store from Surveillance Camera,” *IEEE ISM*, pp.154-159, 2015.
2) D. A. M. Hernandez et al., “How Computer Vision Provides Physical Retail with a Better View on Customers,” *IEEE CBI*, pp.462-471, 2019.
3) J. Candamo et al., “Understanding Transit Scenes: A Survey on Human Behavior Recognition Algorithms,” *IEEE Trans. Intell. Transp. Syst.*, vol.11, no.1, pp.206-224, 2010.